A DIGITAL LIBRARY CONTENT METADATA GENERATOR FOR EPRINTS

AMIR AATIEFF BIN AMIR HUSSIN

A Master's Project submitted in partial fulfilment of the requirements for the degree of Master of Software Engineering

Centre for Graduate Studies

Open University Malaysia

2011

TABLE OF CONTENTS

ABSTRACT

A Digital Library is normally consisting of or made upon a collection of digital objects plus the information and services for storing, accessing and retrieving them. Digital Libraries by nature is a very complex information system. Despite efforts being made to streamline its creation and content population into an out of the box experience, there is still room for automation. For the creation of Digital Library or Online Repositories as it also known, the availability of free open source software such as EPrints developed at University of Southampton, United Kingdom is has simplified the creation process. While the Digital Library software packages such as EPrints have made it easier to create and run Digital Libraries, optimization and customization still needs to be done in order to achieve an optimally usable solution. One the most time consuming tasks involved in setting up a Digital Library is populating these repositories. This can be a very manual task that consumes a large amount of time without automation. One the most time consuming tasks involved in setting up a the content or collections of Digital Library is the data entry that provides detailed information on the available resources which is usually made up of metadata elements that provide information on the content stored. The Digital Library Content Metadata Generator (COMGEN) developed as a part of this project is designed to reduce the workload, time consumption and error prone manual data entry that are being done the traditional way in populating Digital Libraries. COMGEN is built to demonstrate the feasibility of automatic content generation by extracting existing metadata from the source file and transforming it into a usable format for use with the EPrints Import Tool to automatically add new content and populate the Digital Library/Repository.

Keywords: Digital Library, Metadata, EPrints, Generator

ABSTRAK

Perpustakaan Digital lazimnya teridiri daripada satu koleksi objek digital yang mengandungi maklumat, perkhidmatan penyimpanan, penyususan dengan kebolah mengeluarkan semula data serta maklumat tersebut. Perpustakaan Digital secara lazimnya merupakan satu sistem maklumat yang kompleks. Walaupun banyak usaha telah dilakukan untuk penyeragaman dalam pembinaan dan penambahan populasi kandungan Perpustakaan Digital, masih ada ruang untuk automasi. Dalam pembangunan Perpustakaan Digital atau juga dikenali sebagai repositori maya, terdapat perisian 'open source' seperti EPrints yang dibangunkan oleh University of Southampton, United Kingdom yang memudahkan pembangunan sesebuah perpustakaan digital. Selain daripada pakej perisian Perpustakaan Digital seperti EPrints yang telah memudahkan pembangunan dan pengurusannya, penambahbaikan masih perlu dijalankan untuk mendapatkan hasil yang optimum. Penyimpanan dan penambahan koleksi merupakan tugas yang paling sukar dijalankan dalam usaha membangunkan Perpustakaan Digital. Tugas ini boleh dilakukan secara manual tetapi akan mengambil masa yang sangat lama tanpa automasi. Salah satu tugas yang paling lama dalam penyusunan kandungan atau koleksi Perpustakaan Digital ialah memasukkan maklumat yang terperinci daripada sumber dan selalunya dihasilkan daripada elemen metadata yang memberi maklumat tentang kandungan yang disimpan. Digital Library Content Metadata Generator (COMGEN) yang dibangunkan dalam projek ini direka untuk mengurangkan bebanan tugas, megurangkan penggunaan masa dan mengurangkan kesalahan dalam memasukkan data sekiranya dilakukan secara manual. COMGEN dicipta untuk mendemontrasikan keberkesanan penghasilan maklumat secara automatik melalui pengekstrakan metadata yang sedia ada daripada fail sumber dan menukarkannya kepada format yang boleh digunakan dengan ‘Eprints Import Tool’ untuk menambah isi kandungan baru secara automatik ke dalam Perpustakaan/Repositori Digital.

Keywords: Perpustakaan Digital, Metadata, EPrints, Automasi

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview of Project

In a world of rapidly advancing technology and information, many organizations including academic institutions such universities are looking for ways to store digital documents online in order to make them easily accessible and available worldwide. The advent of the Internet or World Wide Web (WWW) has brought to everyone unparalleled amounts of sources knowledge made available through various means such as knowledgebase, online encyclopaedias and an evolution of the original repository of knowledge, the library. The internet today hosts many virtual libraries storing and managing contents in digital form. These repositories commonly referred to as the Digital Library can prove to be a very useful and powerful systems that allows these academic institutions to store, maintain and manage their digital resources. Resources such as documents, collections of thesis and dissertations are stored online making them available and accessible to others as a useful source of references.

According to the Online Dictionary of Library and Information Science (ODLIS) a digital library is "a library in which a significant proportion of the resources are available in digital (machine-readable) format, as opposed to print or microform where the digital content may be locally held or accessed remotely via computer networks" (Reitz, n.d.). This is in direct contrast to traditional libraries that store their collection in print, microform or other media. In simple language Digital Libraries are similar in concept traditional libraries but whereas the traditional library is a physical

building located in a geographical area that contains racks that host thousands of books and takes up a significant amount of space, Digital Libraries consists of information stored electronically in an omnipresent nature that is unlike the method of storing resources within a physical building.

The growth of new Digital Library creations by colleges, universities, associations, and other organizations has created a demand for methods to deal with vast amounts of created or digitized collections of files. There is a need to effectively manage the collection of these digital resources online. Due to the nature of Digital Libraries which are often consisting of multi-format, multi-disciplinary contents, a complete and comprehensive definition for a digital library is difficult.

There is however a more comprehensive definition which can be found from the DELOS Digital Library Reference Model which defines the Digital Library as “An organization, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies” (Candela et al., 2008). This definition highlights the importance of the organizational factor in the Digital Library domain.

At the early stages of its evolution, Digital Library and the term virtual library was initially used interchangeably with one another. This has changed in recent years where virtual library is now primarily used for libraries which aggregate distributed content or virtual in other senses and Digital Library has become the standard term used for libraries storing a centralized digital content repository that can be easily accessible over a local area network or the internet.

Since its inception and rise in everyday use, new methods and ways have been examined with the goal of making the Digital Library creation process easier and

much more efficient. The most common hurdle encountered by organizations when pursuing Digital Library creation is amount of time and resources in planning, development and deployment that are needed in order to implement a successful digital library project. Early Digital Libraries were often custom developed from the ground up consuming time and resources. One solution to this problem was to create pre-built software packages (Gorton, 2007). These software packages help to simplify the process of building, maintaining, managing or running digital libraries.

According to Repository 66, a mash up of worldwide locations indicating open access digital repositories, two most widely used and commonly known software used extensively for Digital Library creations are DSpace developed by the Massachusetts Institute of Technology (MIT) the as a product of the HP-MIT Alliance and EPrints developed at University of Southampton, United Kingdom ("Repository 66", 2010). These software packages for Digital Library creation helps users in creating a basic Digital Library with the ability to accept, store and make accessible information to users without any or maybe very small amount custom programming. These software packages are not without difficulty and unlike common off the shelf software (COTS) used by consumers every day, certain customization and manual work have to be done especially in creating and populating organization or domain specific digital libraries. EPrints will be selected as the Digital Library test bed for this project due to its suitability in adopting automation and its ease of configuration and customization.

While the Digital Library software packages described above have made it easier to create and run Digital Libraries, optimization and customization still needs to be done in order to achieve an optimally usable solution. Populating these repositories can be a very manual task and could consume a large amount of time without automation.

Based on the situation at hand, a software for automatically generating Digital Library content is proposed to automate classification of the digital material contents such documents stored in Portable Document Format (PDF) files into the existing EPrints repository. This will help to speed up the repository population with the elimination of manual data entry of information about the digital content into EPrints. The intent is to automate this task in a way so that minimal or no intervention from the user is required when adding digital content into EPrints. This is especially useful when doing backwards processing of uncategorized or catalogued digital contents that needs to be made available on-line through EPrints in a timely manner.

1.2 Problem Statement

A Digital Library by nature is a very complex information system. Despite efforts being made to streamline its creation into an out of the box experience, there is still room for automation. EPrints is one such effort, a Web and command-line application providing a software package that can be customized to the exact needs and structure of each institution or repository type ("What is EPrints?", n.d.). EPrints is an open source software package designed to help build open access repositories that comply with the Open Archives Initiative Protocol for Metadata Harvesting (“EPrints” 2010). As such EPrints can easily be used as tool for quickly creating an online Digital Library. EPrints repository configuration requires modifying configuration files written in either Perl or XML. On the appearance and user interface side repository look and feel is controlled by HTML templates, CSS style sheets and images.

The standard submission procedure for EPrints contents is normally based on the way EPrints was originally designed to work as, which has traditionally been associated with the so-called self-archiving (Swan, 2005). This is where the authors

themselves format their documents and manually submit them to into the Digital Library themselves. It is normal method used to add new fresh contents to the Digital Library. It does not however address the issue of putting in existing contents that are scattered around or importing contents that currently resides in other repositories or database.

Manual submissions can be very taxing and filling out all the data required by EPrints will be a challenging task when it comes to adding thousands of files into the Digital Library (Carr and Harnad, 2005). Manual work also leaves room for user error and inconsistency in actual data and user input. Automated systems will most likely help to reduce if not totally eliminate this problem. This will in the long run also help to reduce the total operating costs by reducing the number of required personnel needed to manage the Digital Library (Haimson and Grossman, 2009).

The EPrints software has an import facility that supports various formats but this is quite technical in nature and requires a fair amount of IT knowledge and manual configuration. EPrints has a bulk data import facility based on XML but what is lacking is the facility to create the XML specification from existing files such as PDF Documents in an automated fashion ("EPrints Depositing Items" 2007). This is the source of the problem. Submitting documents or contents manually may be feasible if it was a self-submission exercise and the amount of content to be submitted is small.

However when a collection or a set of content containing thousands of files, this process can be tedious and filling up all the required fields manually in EPrints can be a daunting task. To quickly expand EPrints contents into a large full scale Digital Library repository, the Digital Library Content Metadata Generator (COMGEN) is built to facilitate in extracting information from the content to be submitted such as

PDF documents which will be the only file format supported by the COMGEN prototype . Certain information and attributes will be extracted from the PDF file as metadata that will be used for automatically populating an EPrints compliant XML document format which can later be used with the EPrints import facility.

Besides the benefit of reducing errors, personnel requirements and resources by implementing automation ("Automation" 2010), COMGEN will also speed up the accumulation of digital contents by speeding up their addition to the Digital Library through automation. Thus in a shorter timeframe a comprehensive collection can be built into the Digital Library quickly, a process which would have taken a significantly longer time if it was to be completed manually. With this tool, staff could concentrate more on filtering the quality of the contents and watching over other aspects of the system that will directly affect the user experience.

1.3 Objective and Scope

This section outlines the objectives and scope of this project. First the list of objectives for producing COMGEN is outlined followed by the scope of the study.

1.3.1 Objectives

The objective of this research project is to produce a working prototype of COMGEN through the achievement of the defined objectives as follows:

i. Review of the current problem and difficulties related to manual content submission and metadata input into EPrints.
ii. To perform identification of the major obstacles that need to be overcome when extracting metadata, keywords and text harvesting from the digital contents such as PDF files.

iii. To work in producing an overall efficient and functional design of the proposed prototype.

iv. To develop a working prototype for the proposed COMGEN which for evaluation purposes will only work with a single file format, PDF's

v. To conduct tests that will validate the prototype.

1.3.2 Scope of Study

The purpose of COMGEN is to assist in automatically generating contents into a Digital Library based on EPrints. For the purpose of this study, a simple prototype is produced to demonstrate basic functions of automating content generation. It is developed to be a prototype or proof of concept that will only support generating Digital Library contents from a single document format, PDF. Once the generation of content is complete, each of the PDF files along with its corresponding metadata attributes and information will be deposited into an EPrints Test Repository using the EPrints import tool. COMGEN is created to assist in automating the Digital Library population process, to help in rapidly increasing repository size and expansion of contents. The system in its basic form should be able to perform the following functions in order to support the automated content metadata generation process to achieve the main objectives outlined.

i. COMGEN is to provide an interface to import PDF documents metadata extracted using Apache Tika into its input file which includes attributes such subject, author and other metadata extracted from the PDF file.

ii. COMGEN should be able to process metadata information from any PDF files such thesis, dissertations or articles as long as the metadata extracted by Apache Tika has been put into a file recognisable by COMGEN.

iii. COMGEN should be able to use the information extracted and collected to generate XML files compliant with EPrints format to be used with the EPrints import tool.

iv. To be able to export the generated XML file into fully compliant EPrints XML format that can be imported directly into EPrints with minimal intervention using the EPrints import facility.

v. To be able to accept input and process files that will automate XML generation from harvested data.

vi. To be able to support EPrint/Content creation activities into EPrints

1.4 Significance of Study

This project changes the way a Digital Library accumulate its contents and collections. Instead of relying only on user input or manual data entry to build up metadata for each item in the library, an automated system is developed to help in adding items and collections to the digital library quicker and more efficiently. With direct metadata extraction it will also avoid potential problems that might have been caused by human error such as entry of inconsistent data or even wrong data into wrong fields. Besides this, humans also have a limit and work only during a fixed number of hours each day which will also slow down item and collection addition to the Digital Library especially those with a backlog of collections to upload and catalogue. An automated system would eliminate this weakness due to its ability to operate continuously with minimal or no user intervention ("Automation" 2010).

These circumstances are the driving force behind the idea of the COMGEN project. Its benefits among others include faster items and collection build-up into the Digital Library, greatly reduced manpower requirements and most importantly accurate

metadata generation harvested from the items and collections being uploaded. This also ensures the quality and accuracy of the metadata in the Digital Library. If COMGEN is successful in performing all the tasks outlined in the objectives it will revolutionize content loading into Digital Libraries. The outcome will be faster knowledge dispersion and sharing, quicker availability of resources for the users and high quality metadata accumulation.

1.5 Organization of Report

This project report is organized into five chapters. Chapter 1 provides an Overview of project followed by the problems statement which gives information about what this project is attempting to provide solution for. This is followed by objectives and scope after which the significance of study is provided to support the need for this project. It concludes with organization of report which details the organization of information for each chapter.

Chapter 2 is the review of literature which covers other works on Digital Libraries. This is detailed into concept and technologies, metadata, Dublin Core, Digital Library services, and internet services and Digital Libraries. There is also review on Digital Library software which covers two of the most popular software in use today, DSpace and EPrints. Finally the chapter covers automated metadata creation and extraction followed by the conceptual framework.

Chapter 3 covers the research methodology outlining the methodology applied for this research project. This chapter includes the requirement specification, overview of the requirements, scope, and system description, constrains and the specific functional and non-functional requirements of the Project. It also outlines the system design which provides the general overview of the COMGEN design and its architecture.

Next, this chapter outlines the specifications of the input and output files and the implementation and also includes the Test Plan. Chapter 4 provides the result of the testing concluded as outlined in the previous chapter and includes an analysis of the test results. Chapter 5 the final chapter provides the conclusion and a brief description of Future Work.

REFERENCES

ANSI/NISO Z39.85 - The Dublin Core Metadata Element Set (2007) Available from http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key%3Austring%3Aiso-8859-1=9b7bffcd2daeca6198b4ee5a848f9beec2f600e5

Apache Tika - a content analysis toolkit (n.d.) Retrieved 5th November 2010 from http://tika.apache.org/

Apps, A. and Macintyre, R. (2000) Dublin Core Metadata for Electronic Journals - Research and Advanced Technology for Digital Libraries, *Lecture Notes in Computer Science,* 2000, Volume 1923/2000, 93-102, DOI=10.1007/3-540-45268-0_9

Automation (2010) Encyclopedia of Business, 2nd Edition Retreived 30th December 2010 from http://www.referenceforbusiness.com/encyclopedia/Assem-Braz/Automation.html

Brambilla M., Ceri S., Comai S., Dario M., Fraternali P. and Manolescu I. (2004) Declarative specification of Web applications exploiting Web services and workflows. *SIGMOD '04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data,* Paris, France 2004, pp. 909-910 ACM, New York, NY, USA, DOI=10.1145/1007568.1007688

Bray T., Paoli J., Sperberg-McQueen C. M., Maler E. and Yergeau F. (2006) *Extensible Markup Language (XML) 1.0 (Fourth Edition)*. Retrieved 7th October 2010 from http://www.w3.org/TR/2006/REC-xml-20060816/

Candela, L., Castelli, Y., Ioannidis, G., Koutrika, P., Pagano, S., Ross, H., Schek, J. and Schuldt, H. (2006) The Digital Library Manifesto, DELOS: A Network of Excellence on Digital Libraries, Pisa, Italy ISBN: 2-912335-24-8, 2006.

Candela, L.; Castelli, D.; Ferro, N.; Ioannidis, Y.; Koutrika, G.; Meghini, C.; Pagano, P.; Ross, S.; Soergel, D.; Agosti, M.; Dobreva, M.; Katifori, V.; Schuldt, H. (2008).The DELOS Digital Library Reference Model - Foundations for Digital Libraries.Version 0.98 Avaialble from: http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf

Carr, L. and Harnad, S. (2005) Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving. Technical Report, ECS, University of Southampton. (Unpublished Public Draft) Avaialble from: http://eprints.ecs.soton.ac.uk/10688/1/KeystrokeCosting-publicdraft1.pdf

Curbera, F., Duftler F., Khalaf R., Nagy W., Mukhi N., and Weerawarana S. (2003) The next step in Web services. Communications of the ACM - Service-oriented computing, Volume 46 Issue 10, October 2003. ACM, New York, NY, USA, DOI=10.1145/944217.944234

Digital Imaging Tutorial - Metadata.(2010) Retrieved 15th October 2010, from http://www.library.cornell.edu/preservation/tutorial/metadata/metadata-01.html 2003

Duncan, C. and Douglas, P. (2009) Automatic Metadata Generation - Use Cases [White paper electronic version]. Retrieved 5th October 2010 from University of Edinburgh Intrallect website: http://www.intrallect.com/index.php/intrallect/content/download/960/4029/file/synthesis_report.pdf

EPrints Depositing Items (2007). Retrieved 15th October 2010, from http://www.eprints.org/software/training/users/depositing.php

EPrints Handbook – Overview (2010). Retrieved 15 October 2010 from http://www.eprints.org/documentation/handbook/overview.php

EPrints. (2010). Retreived 5th November 2010, from http://www.eprints.org/software/

Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M. (2007) Automated Template-based Metadata Extraction Architecture. Lecture Notes in Computer Science, 2007, Volume 4822/2007, 327-336, DOI: 10.1007/978-3-540-77094-7_42

Galin, D. (2004) Software Quality Assurance – From Theory to Implementation Pearson Publishing. Essex, United Kingdom.

Gorton, D. (2007) *Practical Digital Library Generation into DSpace with the 5S Framework.* (Master's thesis). Virginia Polytechnic Institute and State University, Virginia, USA, 2002.

Haimson, C. and Grossman, J. (2009) A GOMSL analysis of semi-automated data entry. In *Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems (EICS '09)*. ACM, New York, NY, USA, 61-66. DOI=10.1145/1570433.1570445

Leiner, B. M. (1998) The Scope of the Digital Library. *Draft Prepared by Barry M. Leiner for the DLib Working Group on Digital Library Metrics, January 16, 1998.* Revised October 15, 1998. Retrieved 16th October 2010 from http://www.dlib.org/metrics/public/papers/dig-lib-scope.html

Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., Diekema, A., McCracken, N., Silverstein, J. and Sutton, S. (2002) Automatic metadata generation & evaluation. *SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, August 11-15, 2002, Tampere, Finland. New York, NY, USA: ACM Press

Metadata (2010). Retrieved 5th November 2010, from http://www.ukoln.ac.uk/metadata/

Marinai, S. (2009) Metadata Extraction from PDF Papers for Digital Library. Ingest Dipartimento di Sistemi e Informatica - Universit`a di Firenze Via S. Marta, 3 - 50139 - Firenze – Italy 978-0-7695-3725-2/09 2009 IEEE DOI=10.1109/ICDAR.2009.232

Reddy, R. and Wladawsky-Berger, I. (2001) "President's Information Technology Advisory Committee - Panel on Digital Libraries Digital Libraries: Universal Access to Human Knowledge," NationalCoordination Office for Information Technology Research and Development 2001.

Repository 66 (2010) Repository66.org Repository Maps [A mashup indicating the worldwide locations of open access digital repositories]. Repository66.org Project based on data provided by ROAR and the OpenDOAR service developed SHERPA (Securing a Hybrid Environment for Research Preservation and Access) Retrieved from: http://maps.repository66.org/

Reitz, J.M. (n.d.) *Online Dictionary for Library and Information Science.* Retrieved from http://www.abc-clio.com/ODLIS/odlis_d.aspx#digitallibrary

Setting up an institutional e-print archive (2010) Retrieved 21 October 2010 from http://www.ariadne.ac.uk/issue31/eprint-archives/

Smith M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R. and Walker J. H. (2003) DSpace — An Open Source Dynamic Digital Repository, D-Lib Magazine January 2003 Volume 9 Number 1 ISSN 1082-9873 Retrieved from http://www.dlib.org/dlib/january03/smith/01smith.html

Sommerville, I. (2007) Software Engineering 8th Edition Pearson Publishing. Essex, United Kingdom.

Sperberg-McQueen C. M. and Thompson, H. (2000) *W3C XML Schema*. Retrieved 7th October 2010, from http://www.w3.org/XML/Schema

Suleman H. and Fox E. A. (2002) Designing protocols in support of digital library componentization, in *Research and Advanced Technology for Digital Libraries.6th European Conference*, ECDL 2002. Proceedings, 16-18 Sept. 2002, Rome, Italy, 2002, pp. 568-82 ISBN - 3-540-44178-6.

Swan, A. (2005) Open access self-archiving: An Introduction. Technical Report, JISC, HEFCE. Retreived 9th December 2010 from http://www.jisc.ac.uk/uploaded_documents/JISC-BP-OpenAccess-v1-final.pdf

Taylor, C. (2003) An Introduction to Metadata, a paper written by Chris Taylor - Manager, Information Access Service, University of Queensland Library. Retrieved from http://www.library.uq.edu.au/iad/ctmeta4.html

The Dublin Core Metadata Initiative (2010) Retrieved 7th October 2010 from http://dublincore.org/

What is EPrints? (n.d.) Retrieved 9th November 2010 from the EPrints Wiki http://wiki.eprints.org/w/Introduction